Combining and splitting bunsetsu of the Kyoto Text Corpus

Zhen ZHOU* Alastair BUTLER†‡ Kei YOSHIMOTO*‡
*Graduate School of International Cultural Studies, Tohoku University
†PRESTO, Japan Science and Technology Agency
‡Center for the Advancement of Higher Education, Tohoku University

syusin3@yahoo.co.jp

Abstract
The Kyoto Text Corpus contains a wealth of syntactic information. Accessing this information is however non-trivial, e.g., to build meaning representations with appropriately scoped material. In this paper we discuss an automated modification that may combine or split apart bunsetsu. The one constraint obeyed is that changes should not disrupt the overall coding of bunsetsu dependencies. Combining (e.g., uniting compound functional expressions) involves the placement of content from one bunsetsu into a prior bunsetsu but without any removal of dependency information. Splitting (e.g., to allow clause level scope placements for compound auxiliary verbs, negation and sentence final particles) involves the creation of new bunsetsu as fractions of the split bunsetsu. We illustrate gains made in the quality of semantic representations we can produce from the Kyoto Text Corpus over what was possible without combining and splitting.

1 Introduction

The Kyoto Text Corpus (Kurohashi and Nagao, 2003) is a morphologically and syntactically annotated corpus for 40,000 sentences of Mainichi Shimbun newspaper articles for 1995. In addition with version 4.0 of the corpus (Kawahara et al., 2005) a subset of 5,000 sentences are annotated with case, anaphora and coreference information. This offers a considerable wealth of gold standard syntactic information. However taking this information as a basis for building meaning representations with appropriately scoped material is non-trivial. This paper discusses an automated modification of the Kyoto Text Corpus that may combine or split apart bunsetsu to better serve as parsed information that can thereafter be used to reach meaning representations for sentences.

The paper is organised as follows. Section 2 motivates combining bunsetsu. Section 3 motivates splitting bunsetsu. Section 4 details our changes to the Kyoto Text Corpus. Section 5 notes supplementary information added when changes are made. Section 6 concludes.

2 Motivation for combining bunsetsu

Japanese has many expressions formed from multiple morphemes that can have a non-literal functional usage. NINJAL (2001) detail 125 major expressions with variants to total 337 expressions. In this section we illustrate problems functional expressions cause for further processing of the Kyoto Text Corpus, focusing on について.

In (1a) particle に has the case-marking function of 'with' modifying the verb ついて 'keep close contact' to produce the literal content meaning of 'follow'. By contrast, in (1b) について has a case-marking function as a single unit similar to the English preposition 'about'.

(1) a. 私は彼について走った。
I ran following him.

b. 私は彼について話した。
I talked about him.

Following the annotation scheme of the Kyoto Text Corpus (1a) can be analysed as in (2).

(2)
```
# S-ID:2
* 0 3D
+ 0
私 わたし * 名詞 普通名詞 * *
は は * 助詞 副助詞 * *
* 1 2D
+ 1
彼 かれ * 名詞 普通名詞 * *
に に * 助詞 格助詞 * *
* 2 3D
+ 2 <rel type="ガ" tag="0"/><rel type="に" tag="1"/>
ついて ついて つく 動詞 * 子音動詞カ行 タ系連用テ形
* 3 -1D
+ 3 <rel type="ガ" tag="0"/>
走った はしった 走る 動詞 * 子音動詞ラ行 タ形
。 。 * 特殊 句点 * *
EOS
```

The first line of (2) is the sentence ID. Lines starting with a star begin bunsetsu, with numbers to specify the current bunsetsu ID and the ID of the target bunsetsu for the dependency. The character after the target bunsetsu ID specifies the type of dependency, notably D for direct dependency. A bunsetsu without a target bunsetsu is the root of the sentence, and this is indicated by "`-1D`". Lines starting with a plus sign code relation information, with minimally an ID number. Relation information can be the target for other relation information—notably for case, anaphora and coreference—and this is coded with `rel` tags containing the attribute `type` to indicate the kind of relation, e.g., が marks a grammatical subject, and `tag` to state the relation information ID of the modifier element.

Sufficient structural and case information can be obtained from (2) as a basis for assembling the meaning representation of (3). Representation (3) assumes a Davidsonian theory (Davidson, 1967) for coding verbs as predicates that have minimally an implicit event argument. The verbs of (3) also have subject arguments. Furthermore events are existentially quantified over and may be further constrained, as is the case with e_1 which is coded to be an event that occurs with (に) some value that is further restricted to be 彼 'him'.

(3) $\exists e_1 e_2 x(\text{彼}(x) \wedge \text{つく}(e_1,\text{私}) \wedge \text{に}(e_1) = x \wedge \text{走る}(e_2,\text{私}))$

Similarly, following the Kyoto Text Corpus annotation scheme, we can expect (1b) to receive the bunsetsu dependency analysis of (4). Aside from the different main verb, (4) differs from (2) only with respect to the case information: ついて no longer has case information, while 話した is coded to receive a について argument. Notably (4) and (2) share the representation of bunsetsu dependencies in having について span two bunsetsu.

(4)

```
# S-ID:4
* 0 3D
+ 0
私 わたし * 名詞 普通名詞 * *
は は * 助詞 副助詞 * *
* 1 2D
+ 1
彼 かれ * 名詞 普通名詞 * *
に に * 助詞 格助詞 * *
* 2 3D
+ 2
ついて ついて つく 動詞 * 子音動詞カ行 タ系連用テ形
* 3 -1D
+ 3 <rel type="ガ" tag="0"/><rel
                            type="について" tag="1"/>
話した はなした 話す 動詞 * 子音動詞サ行 タ形
。 。 * 特殊 句点 * *
EOS
```

A meaning representation for (1) is given in (5). The verb 話す 'talked' expresses a predicate taking as arguments the subject 私 'I' and an event that is further constrained to be about (について) 彼 'him'. This clearly exhibits the functional role of について. Our problem is that reading this from (4) is complicated by に and ついて spanning two bunsetsu.

(5) $\exists e_1 x(\text{彼}(x) \wedge \text{話す}(e_1,\text{私}) \wedge \text{について}(e_1) = x)$

Adjustment of the bunsetsu dependency representation to (6) offers a representation that more readily maps to (5).

(6)

```
# S-ID:6
* 0 3D
+ 0
私 わたし * 名詞 普通名詞 * *
は は * 助詞 副助詞 * *
* 1 3D
+ 1
彼 かれ * 名詞 普通名詞 * *
について について * binding 格助詞 * *
* 2 3D
+ 2
* 3 -1D
+ 3 <rel type="ガ" tag="0"/><rel
                            type="について" tag="1"/>
話した はなした 話す 動詞 * 子音動詞サ行 タ形
。 。 * 特殊 句点 * *
EOS
```

Note adjustment to (6) occurs without removal of dependency information (lines starting with a star) or of relation information (lines starting with a plus sign), since any disruption to this information has potential to break the `rel` information that codes case, anaphora and coreference information. Also note that the change from (4) to (6) can be automated without fear of erroneously adjusting (2), since the changes can be made dependent on the presence of case information, specifically `<rel type="ニツイテ" tag="1">` with 話した.

3 Motivation for splitting bunsetsu

To ease building meaning representations, we wish elements that form scopal operations to be captured as distinct 'bunsetsu'. However ない 'not' is analysed in the Kyoto Text Corpus as part of a larger bunsetsu containing a predicate e.g., (7) will be annotated as (8).

(7) 洋子はコンピューターを買えない。
Yoko is unable to buy a computer.

(8)

```
# S-ID:8
* 0 2D
+ 0
洋子 ひろこ * 名詞 人名 * *
は は * 助詞 副助詞 * *
* 1 2D
+ 1
コンピューター こんぴゅーたー * 名詞 普通名詞 * *
を を * 助詞 格助詞 * *
* 2 -1D
+ 2 <rel type="ガ" tag="0"/><rel type="ヲ" tag="1"/>
買え かえ 買える 動詞 * 母音動詞 未然形
ない ない * 接尾辞 形容詞性述語接尾辞 イ形容詞アウオ段 基本形
。 。 * 特殊 句点 * *
EOS
```

With (8) as a basis for building a meaning representation the scope of negation is restricted to scope only over the verb, to for example derive the meaning representation of (9).

(9) $\exists e_1 x(\text{コンピューター}(x) \wedge \text{ない}(\text{買える}(e_1,\text{洋子},x)))$

For (9) to be true there should be some computer and some event such that the event is not Yoko buying the computer. By contrast (7) will be true if there are no computers. This is captured by (10) where negation scopes over the quantification of computer and event.

(10) $\text{ない}(\exists e_1 x(\text{コンピューター}(x) \wedge \text{買える}(e_1,\text{洋子},x)))$

To derive (10) we need an analysis along the lines of (11), with ない as a distinct 'bunsetsu' that is also the root of the sentence.

(11)

```
# S-ID:11
* 0 2D
+ 0
洋子 ひろこ * 名詞 人名 * *
は は * 助詞 副助詞 * *
* 1 2D
+ 1
コンピューター こんぴゅーたー * 名詞 普通名詞 * *
を を * 助詞 格助詞 * *
* 2 2.5D
+ 2 <rel type="ガ" tag="0"/><rel type="ヲ" tag="1"/>
買え かえ 買える 動詞 * 母音動詞 未然形
* 2.5 -1D
ない ない * operation 形容詞性述語接尾辞 イ形容詞アウオ段 基
本形
。 。 * 特殊 句点 * *
EOS
```

Having ない as a distinct 'bunsetsu' is achieved in (11) by creating a new bunsetsu that is a fraction of the original 買えない bunsetsu. Keeping to adding only fractions of bunsetsu ensures there is no disruption to the bunsetsu dependency analysis for the sentence as a whole.

4 Changes to the corpus

This section details the changes we make to the bunsetsu structure of the Kyoto Text Corpus. Changes are made automatically as part of a pipeline for taking treebank data as input and generating meaning representations as output. Consequently changes are applied with each run of the system, with the original Kyoto Text Corpus data remaining unaltered.

4.1 Combining particle functional expressions

Particle functional expressions formed from multiple morphemes comprise two adjacent bunsetsu, with the prior bunsetsu dependent on the latter. The prior bunsetsu contains one or two particles, while the latter contains either a verb or a verb and a particle. We currently combine 96 cases of functional expressions. Combining occurs when:

1. the current bunsetsu has X and next bunsetsu has Y, e.g., X=とand Y=してcombining as として.
2. the current bunsetsu has X and next bunsetsu has Y and Z, e.g., X=に, Y=よる and Z=と combining as によると.
3. the current bunsetsu has X and Y and next bunsetsu has Z, e.g., X=から, Y=と and Z=いって combining as からといって.

Because combining occurs automatically we have to look out for clues to warrant the combining. In this regard when available case information is very reliable, as demonstrated in section 2 with について. Unfortunately there is not always case information to correspond to the functional use of a compound particle expression. Tuning the system to respond to other clues, such as the wider bunsetsu structure and aspects of bunsetsu content is work in progress.

4.2 Combining suffix functional expressions

In the Kyoto Text Corpus suffix compound functional expressions have two bunsetsu structures. One is where every morpheme is included in the same bunsetsu. Another type is that morphemes are split into two adjacent bunsetsu. Combining occurs when:

1. the current bunsetsu has X and Y, e.g., X=一方 and Y=だ combining as 一方だ.
2. the current bunsetsu has X and Y and Z, e.g., X=なければ, Y=なら and Z=ない combining as なければならない.
3. the current bunsetsu has X, Y, Z, E, e.g., X=ざる, Y=を, Z=得 and E=ない combining as ざるを得ない.
4. the current bunsetsu has X, Y, Z, E, F, G, e.g., X=ず, Y=に, Z=は, E=い, F=られ and G=ない combining as ずにはいられない.
5. the current bunsetsu has X and next bunsetsu has Y, e.g., X=に and Y=違いない combining as に違いない.
6. the current bunsetsu has X and next bunsetsu has Y and Z, e.g., X=に, Y=すぎ and Z=ない combining as にすぎない.
7. the current bunsetsu has X and Y and next bunsetsu has Z, e.g., X=こと, Y=が and Z=ある combining as ことがある.

8. the current bunsetsu has X and Y and next bunsetsu has Z and E, e.g., X=わけに, Y=は, Z=いか and E=ない combining as わけにはいかない.

Altogether 29 cases of compound functional expressions are combined.

4.3 Combining other functional expressions

The functional expressions we have combined are based on the corpus work of Tsuchiya et al. (2005). In addition we see benefits when combining other expressions, for example, expressions following the schema of のXに, where X is e.g., 上 'on'. As an example, consider (12).

(12) 机の上にねこがいます。
There is a cat on the desk.

Without combining we generate the meaning representation of (13a), while combining allows for (13b).

(13) a. $\exists e_1 xyz(\text{机}(x) \wedge \text{の\_上}(z,x) \wedge \text{ねこ}(y) \wedge \text{いる}(e_1,y) \wedge \text{ニ}(e_1) = z)$

b. $\exists e_1 xy(\text{机}(y) \wedge \text{ねこ}(x) \wedge \text{いる}(e_1,x) \wedge \text{の上に}(e_1) = y)$

4.4 Splitting

In addition to ない 'not', discussed in section 3, we have found splitting bunsetsu information to be necessary in the case of processing suffix functional expressions such as なければならない 'must' that will serve as scopal operations in a meaning representation, as well as sentence final particles, e.g., the question operator か.

5 Supplementary information

While splitting and combining we also add tag information useful for subsequent processing. Split suffix functional expressions, sentence final particles and negation are tagged as `operation` so as to lead to a scopal operation, as seen with ない in (11). Combined particle functional expressions are classified into three categories:

1. Those subsequent to a nominal, which mainly function as case-marking particles, are tagged as `binding`, e.g., について in (6).

2. Those subsequent to a predicate, which mainly function as conjunctive particles, are tagged as `coord`, e.g., となると.

3. Those subsequent to a nominal, which mainly function as adnominal particle, are tagged as `embed`, e.g., という.

6 Conclusion

In this paper we presented techniques for combining and splitting bunsetsu that are part of a bunsetsu dependency analysis. Combining involves the placement of content from one bunsetsu into a prior bunsetsu. Combining bunsetsu information is necessary when the bunsetsu involve compound expressions that have a functional role. Splitting involves the creation of new bunsetsu as fractions of the split bunsetsu. Splitting is necessary in the case of elements that will serve as scopal operations. We described carrying out splitting and combining of bunsetsu automatically on the content of the Kyoto Text Corpus, but in such a way as to avoid changes that could disrupt the overall coding of bunsetsu dependencies and relation information. Finally we were able to add structural information—`binding`, `coord` and `embed`—as tags for combined and split bunsetsu as an aid for subsequent processing.

This research has been supported by the JST PRESTO program (Synthesis of Knowledge for Information Oriented Society).

References

Davidson, Donald. 1967. The logical form of action sentences. In N. Rescher, ed., *The Logic of Decision and Action*. Pittsburgh: University of Pittsburgh Press. Reprinted in: D. Davidson, 1980. *Essays on Actions and Events*. Claredon Press, Oxford, pages 105–122.

Hashimoto, Shinkichi. 1934. *Essentials of Japanese Grammar (Kokugoho Yousetsu)*. Iwanami. (In Japanese).

Kawahara, Daisuke, Ryohei Sasano, Sadao Kurohashi, and Koichi Hashida. 2005. Specification for annotating case, ellipsis and coreference. Kyoto Text Corpus Version 4.0. (In Japanese).

Kurohashi, Sadao and Makoto Nagao. 2003. Building a Japanese parsed corpus – while improving the parsing system. In A. Abeillé, ed., *Treebanks: Building and Using Parsed Corpora*, chap. 14, pages 249–260. Dordrecht, The Netherlands: Kluwer Academic Publishers.

NINJAL. 2001. Gendaigo hukugouji youreishu. Tech. rep., National Institute for Japanese Language and Linguistics. in Japanese.

Tsuchiya, Masatoshi, Takehito Utsuro, Suguru Matsuyoshi, Satoshi Sato, and Seiichi Nakagawa. 2005. A corpus for classifying usages of Japanese compound functional expressions. In *Proceedings of the Pacific Association for Computational Linguistics*, pages 345–350.